Nascimento et al. BMC Proceedings 2011, 5(Suppl 7):P170
http://www.biomedcentral.com/1753-6561/5/S7/P170

BMC Proceedings

POSTER PRESENTATION Open Access

An integrated database of *Eucalyptus* spp. genome project

Leandro Costa Nascimento[1*], Jorge Lepikson Neto[1], Marcela Mendes Salaza[1], Eduardo Leal Oliveira Camargo[1], Wesley Leoricy Marques[1], Danieli Cristina Gonçalves[1], Ramon Oliveira Vidal[2], Gonçalo Amarante Guimarães Pereira[1], Marcelo Falsarella Carazzolle[3]

From IUFRO Tree Biotechnology Conference 2011: From Genomes to Integration and Delivery
Arraial d'Ajuda, Bahia, Brazil. 26 June - 2 July 2011

Background

The species of the genus *Eucalyptus* are the most planted for the fiber crop in the world. They are mainly utilized for timber, pulp and paper production. Brazil, helped by the favorable weather conditions, appears as a big producer and exporter of eucalyptus derivates. In 2002, the Brazilian network research of the *Eucalyptus* Genome (Genolyptus) was established with the goal of integrating several academic and private institutions currently working with eucalyptus genomics in Brazil. This project generated around 200.000 ESTs from several tissues and conditions. Consequently, several individual projects have been implemented generating other transcriptome databases, in special, using RNA-Seq technology. In 2010, a draft genome (http://eucalyptusdb.bi.up.ac.za) of the specie *E. grandis* was produced by researches of the Joint Genome Institute (DOE-JGI) and the *Eucalyptus* Genome Network (EUCAGEN). The main goal of this work is to develop an *Eucalyptus*database (http://www.lge.ibi.unicamp.br/genolyptus) integrating public and private data in a friendly and secure web interface with bioinformatics tools that allowing the users perform complex searches.

Results and discussion

First, the public and private ESTs (130,290 from Genolyptus and 36,981 from NCBI) were assembled producing 48,760 unigenes (17,795 contigs and 30,765 singlets). Basically, the bdtrimmer [1] and CAP3 [2] programs were used to perform sequence trimming (exclude vector, ribosomal, low quality and too short reads) and sequence assembly, respectively.

The autofact pipeline [3] was used to perform an automatic annotation of the assembled unigenes based on BLAST [4] searches, e-value cutoff of 1e-5, against some protein databases, including: non-redundant (NR) database of NCBI, uniref90 and uniref100 – databases containing only curated proteins [5], pfam – database of proteins families [6], kegg – database of metabolic pathways [7] and Gene Ontology (GO) – database of functional annotation [8].

The Genomic and Expression Laboratory at State University of Campinas (http://www.lge.ibi.unicamp.br) sequenced ten RNA-Seq libraries from four species (*E. Urograndis*, *E. globulus*, *E. grandis* and *E. urophylla*) using the Illumina/Solexa technology. Additionally, three RNA-seq libraries [9] were downloaded from NCBI (SRA – sequence read archive). All RNA-seq reads were aligned against the assembled unigenes and genome assembly using the SOAP2 [10] and TopHat [11] aligners, configured to allow up two mismatches, discard sequences with "N"s and return all optimal alignments.

In order to perform a differential expression analysis between ESTs or RNA-seq libraries some normalization pipelines and statistical tests have been implemented. From ESTs, the differentially expressed genes between libraries were performed applying AC test [12] in assembled unigenes. The results are available to the users by a web interface (called Electronic Northern) that allows searches by gene or library name. Furthermore, it is possible to compare the gene expression between two or more libraries. From the RNA-seq libraries, the DEG-seq software [13] was used to perform normalization and statistical analysis considering 99% of confidence rate (cut-off of 0.01).

* Correspondence: leandro@lge.ibi.unicamp.br
[1]Laboratório de Genômica e Expressão - Instituto de Biologia - Universidade Estadual de Campinas – UNICAMP, Brazil
Full list of author information is available at the end of the article

Figure 1 Home-page of the *Eucalyptus* database, hosted at http://www.lge.ibi.unicamp.br/genolyptus.

To integrate all data described above, we developed a web site (Fig. 1) hosted in a Fedora Linux machine with MySQL database server. The web interface is based on a combination of CGI scripts using PERL language (including BioPerl module) and the Apache Web Server. The site contains many bioinformatics tools allowing the user perform keyword or local BLAST search in assembled unigenes. Also it is possible to connect these results with gene expression analysis. Moreover, the Gbrowse software (Generic Genome Browser) (Fig. 2)

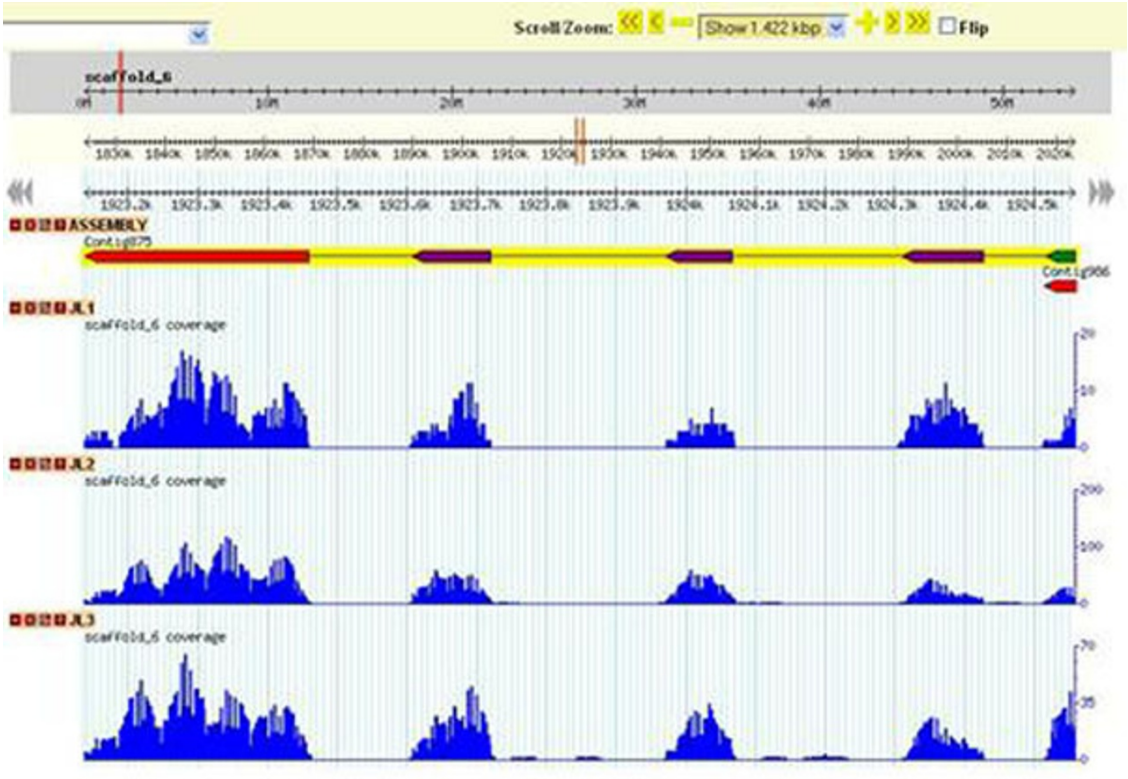

Figure 2 Gbrowse interface of the *Eucalyptus* database. Using Gbrowse is possible to compare gene expression between the RNA-Seq libraries.

was used to visualization the data in a genomic context, integrating the different information by clickable tracks. The top track is the reference genome assembly and the other tracks correspond to assembled unigenes and RNA-seq data mapped into reference.

Acknowledgments

The authors would like to acknowledge all researches of the Joint Genome Institute (DOE-JGI) and the *Eucalyptus* Genome Network (EUCAGEN), responsible to produce the draft genome of the *E. grandis*. Moreover, we thank the CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil) for the financial support of this work.

Author details

[1]Laboratório de Genômica e Expressão - Instituto de Biologia - Universidade Estadual de Campinas – UNICAMP, Brazil. [2]Laboratório de Genômica e Expressão - Instituto de Biologia - Universidade Estadual de Campinas - UNICAMP/LNBio - Laboratório Nacional de Biociências – ABTLuS, Brazil. [3]Laboratório de Genômica e Expressão - Instituto de Biologia - Universidade Estadual de Campinas - UNICAMP/Centro Nacional de Processamento de Alto Desempenho em São Paulo, Universidade Estadual de Campinas-UNICAMP, Brazil.

Published: 13 September 2011

References

1. Baudet C, Dias Z: **New EST Trimming Strategy.** In *Brazilian Symposium on Bioinformatics. Volume 3594.* Lecture Notes in Bioinformatics – Berlin – Germany: Springer – Verlag; 2005:206-209.
2. Huang X, Madan A: **CAP3: A DNA Sequence Assembly Program.** *Genome Res* 1999, **9**:868-877.
3. Koski LB, Gray LW, Lang BF, Burger G: **AutoFACT: An Automatic Functional Annotation and Classification Tool.** *BMC Bioinformatics* 2005, **6**:151.
4. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucl. Acids Res* 1997, **25(17)**:3389-3402.
5. Suzek BE, Huang H, McGarvey P, Mazumber R, Wu CH: **Uniref: comprehensive and non-redundant UniProt reference clusters.** *Bioinformatics* 2007, **23(10)**:1282-1288.
6. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL: **The Pfam Protein Families Database.** *Nucl. Acids Res* 2002, **30(1)**:276-280.
7. Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucl. Acids Res* 2000, **28(1)**:27-30.
8. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 2000, 2525-29.
9. Mizrachi E, Hefer CA, Ranik M, Joubert F, Myburg AA: **De novoassembled expressed gene catalogue f a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq.** *BMC Genomics* 2010, **11(681)**:1471-2164.
10. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment.** *Bioinformatics* 2009, **25(15)**:1966-1967.
11. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25(9)**:1105-1111.
12. Audic S, Claverie JM: **The significance of Digital Gene Expression Profiles.** *Genome Res* 1997, **7**:986-995.
13. Wang L, Feng Z, Wang X, Wang X, Zhang X: **DEGSeq: an R package for identifying differentially expressed genes from RNA-seq data.** *Bioinformatics* 2010, **26(1)**:136-138.

doi:10.1186/1753-6561-5-S7-P170
Cite this article as: Nascimento *et al.*: An integrated database of *Eucalyptus* spp. genome project. *BMC Proceedings* 2011 **5**(Suppl 7):P170.